Google Slide Speaker Notes:

**Slide 1 – Title Page**

“Instacart: A Market Basket Analysis with Customer and Product Segmentation”

**Slide 2 – “What Is Instacart”?**

“Instacart is an American company currently operating a grocery delivery service in the United States and Canada. It allows customers to order groceries from participating retailers via the Instacart app with the shopping being done and delivered to the customers’ door by a personal shopper.”

**Slide 3 – Instacart CEO Quote**

“In a matter of a couple of weeks, we were already ahead of our end-of-year goal. A week later, we were ahead of our 2021 goals, and a few days after that, we were ahead of our 2022 goals.”

“Instacart CEO, Apoorva Mehta, is quoted here referencing the soaring Instacart activity in the early stages of COVID-19. As early as February 2020, Instacart reported noticing an increase in orders for home essentials such as toilet paper. During the subsequent lockdown, Instacart became an essential service for millions of Americans trapped at home. Even now as COVID restrictions are being lifted, analysts expect sustained growth in this field as Instacart users have discovered the ease of ordering online and now prefer it to the traditional grocery shopping experience. This prediction is unsurprising with Instacart boasting an estimated 9.6 million active users, over 500,000 delivery shoppers and in March 2021, the company being valued at $39 billion.”

**Slide 4 – Our Selected Topic**

Researching and contemplating the ideal project subject proved to be a monumental task. Our group floated from one idea to the next until we came across an Instacart dataset on Kaggle. Upon initial overview of the dataset, it contained ample information including seven corresponding csv files with data on orders and prior orders and files containing categorization assistance like aisles and departments for available and sold products. This dataset is comprised of about 1.3 million grocery orders placed by around 200,000 customers. With agreement, our group moved forward with the idea of customer and product segmentation along with a market basket analysis. The goal is to determine purchasing patterns among these products as well as gain insight on customer segmentation.

**Slide 5 – Exploratory Phase Intro Slide**

**Slide 6 – Exploratory Data Facts**

* Data source found on kaggle in as various csv files
* Some clean-up was required before uploading to db
* Customer dataset - Leora’s notes
* Orders/products datasets - check which columns have null values, verify all datatypes match what was required, and limit dataset size to make working with datasets more feasible
* **Slide 7 – The Exploratory Phase**
* Data source from kaggle in various csv files
* Used common columns to create related tables
* PostgreSQL database connected to AWS RDS instance for connectivity
* Customer dataset - Leora’s notes
* Orders\products dataset - check which columns have nulls verify all datatypes match desired
* Created orders by product and reorders by columns tables for machine learning

**Slide 8 – Questions We Hope to Answer**

The customer segmentation questions our group hopes to answer with the Instacart data are:

* What is the distribution of annual Income by age?
* What is the distribution of gender for Instacart users?
* What is the age group of most frequent Instacart users?
* What is the overall distribution of income level for instacart users?
* How do male and female spending scores differ?

**Slide 9 – Questions We Hope to Answer**

* What are the Top 10 most ordered products?
* What are the Top 10 most reordered products?
* Which 2 products were ordered together the most?
* Which day and time of the week was the busiest in terms of number of orders received?
* Which products were not frequently reordered? This could help determine which products did not meet customers expectations.
* What were the top aisles with the least amount of orders? This could be a recommendation that we make to our vendors in terms of restructuring their aisles.

**Slide 10 – Analysis Phase Intro Slide**

* Machine learning introduction
* Tableau connection links

**Slide 11 – Graph**

**Slide 12 – Graph**

**Slide 13 – Graph**

**Slide 14 – Machine Learning – Customer Segmentation**

The first instance of Machine Learning used in our project is for customer segmentation. Using KMeans and clustering the Instacart customer data, we will try to successfully create four main categorization groups: Target, Standard, Careless, and Careful. These refer to the ratio of income level and the customer spending score.

**Slide 15 - DB Scan – KNN**

The DBSCAN, or Density Based Clustering. The DBSCAN algorithm groups clusters as areas of high density separated by areas of low density. Because dbscans aren’t grouping in elliptical shapes around a centroid of the clusters. Our DBSAN identified several clusters and these clusters are useful if we were interested in creating a more tailored promotion initiatives since these consumers are very close in their habits towards spending.

(Density = Number of points within a specified radius

DBSCAN uses two different parameters:

Epsilon: Determines a specified radius that if includes enough number of points within, we call it dense area.

Minimum Samples: Determines the minimum number of data points we want in a neighborhood to define a cluster.)

The KNN model.

The knn model relies on supervised data, [K-nearest neighbors](http://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm) is a classification (or regression) algorithm that in order to determine the classification of a point, combines the classification of the K nearest points. It is supervised because you are trying to classify a point based on the known classification of other points. (In addition a number of k values had to be tried before we reached our desired results) Our group favored the KMeans algorithm. (Lazy algorithm) The centroid here is random and you have to determine the number of clusters. We tried 3, 4, 5.

**Slide 16 – K-Means**

The KMEANS model proved most useful for our goal of creating a few clusters and to categorize them and rank them based on their relationship to income and spending.

We have segmented our customers As follows:

Careless, Standard, Target, Careful.

Careless, (light pink Annual income are low but the spending is high.

Standard, (cyan)The annual income and spending score both are in mid range.

Target, (Dark Blue). Annual income and spending score are higher.

Careful, (magenta) Annual income is high but spending is low.

Grouping our customers based on their score and income will help us determine how to target specific demographics and engage them through a variety of promotions.

**Slide 17 – Dendrogram**

**Hierarchical dendrograms are really useful because it shows all the possible linkages within a cluster.**

We used an agglomerative algorithm which takes a little while to run. The way it works, is it starts from the bottom up, meaning it starts with all the various individual data samples, and then groups them into clusters based on the their similarities (euclidian Distance) until you are left with one final cluster. This is great and useful because it illustrates the inner workings of the methodology used to cluster and their relationship. (We may have overlooked this area …many subsets)

**Slide 18 – Machine Learning Product Segmentation**

With almost 50,000 products in our dataset, studying them individually or even based on aisles or departments can prove to be very challenging. Hence, analyzing them in clusters or groups would be beneficial to boost sales, offer promotions and increase overall revenue.  For this purpose, we combined the products dataset with the prior orders dataset to get metrics such as **num\_of\_orders** and **num\_of\_reorders** per product. Also using product characteristics such as its department and the aisle it belongs to, we used K-Means algorithm to cluster the products into 4 main clusters –

· Class 0 - never re-ordered.

·   Class 1 - occasionally re-ordered.

·   Class 2 - often re-ordered or **Popular.**

·   Class 3 - frequently re-ordered or **“In-High Demand”.**

**Slide 19 – Product Clustering**

Using Principal Component Analysis, the products in the “product\_id” column have been combined into the three key class categorizations in accordance with the “num\_of\_reorders” column.

Possible switch to notebook

**Slide 20 – Neural Networks**

Will this product stocked be ever re-ordered? Stores often carry products that are not as frequently bought as other daily used products such as milk, fruits, etc. Predicting the likelihood that a particular product will be re-ordered is necessary for stores for inventory management. This will also be helpful in decision making for the introduction of new products. Machine learning algorithms such as deep learning neural networks can be employed for this purpose since our data is complex and enormous in size. We used the prior orders dataset and joined with products, departments, and aisles to build our combined dataset used to train the Neural Network model.

Features:

·   order day of week (order\_dow)

·   order hour of day (order\_hour\_of\_day)

·   days since product was bought since prior order (days\_since\_prior\_order)

·   order in which product was added to the cart (add\_to\_cart\_order)

·   total number of re-orders for the product (num\_of\_reorders)

·   total number of orders for the product (num\_of\_orders)

·   product department (department)

·   product aisle (aisle)

Target = “**reordered**”

**Slide 21 – Training the Model**

We used pd.dummies for string value features and scaled the dataset before training the model.

After multiple attempts, altering the number of inputs, activation functions and the number of hidden layers and the number of epochs, we were able to achieve an accuracy of 70.5%

**Slide 22 – Results Intro Slide**

**Slide 23 – Customer Segmentation Results**

**Slide 24 – Graph**

**Slide 25 – Graph**

**Slide 26 – Product Segmentation Results**

**Slide 27 – Conclusion Slide**

**Slide 28 – Technologies Used**

**Slide 29 – Recommendations for Future Analysis**

**Slide 30 – What We May Have Done Differently**